LSP International Journal, Vol. 1, June 2014, 41–50

Lexical Features of Academic Writing

Aliakbar Imani & Hadina Habil
Language Academy, Universiti Teknologi Malaysia, Johor Bahru

ABSTRACT

In measuring the quality of written text, especially academic writing, lexical features are as important as grammatical features and should not be ignored. The highly computable nature of lexicons can make them a good criterion for determining and measuring the quality of text. In this article three lexical features: lexical density, complexity, and formality are reviewed and justified as measurement tools of academic texts. Furthermore, a measurement method is offered to evaluate lexical complexity level of an academic text.

1.0 INTRODUCTION

Among various English as a second language writing skills, academic writing has its own significant place in the contemporary research. This is firstly due to the increasing number of international students studying overseas, and secondly due to the important role of academic writing to graduate students in their efforts to be members of their academic community. Textual features of text, in general, can be categorized as discourse features (Swales 1990, Samraj 2002, 2005, and 2008) as well as lexico-grammatical features (Hyland 2008a, 2008b, Wei and Lei, 2011). Although the term lexico-grammar offered by Halliday (1975) truly reveals the inseparability, interrelatedness and interdependency of grammar and vocabulary – as in lexical bundles which are defined as combinations of words that occur repeatedly with a fairly high frequency in a given register (Biber *et al*., 1999: 992) – grammatical and lexical features can be obviously studied separately, since in this sense they can reflect two various aspects of language mastery. It is in this research trend that by a glance at the literature one can easily discover that more attention has been given to grammatical features (Billig 2008, Crawford 2005, Hinkel 2004b, 2002, 2001, Master 1991), than lexical features (Gregg *et al*. 2002, Coxhead 2000).

The first reason of paying more attention to grammatical features than lexical features might be that grammatical studies are more organized and of a longer history, dating back to the initiation of language teaching and learning studies. This has led to the publication of various grammar books and guidelines as well as the introduction of various second language (L2) teaching methodologies which are

**Correspondence to:* Hadina Habil (email: hadina@utm.my)

either completely or partly based on grammar. The position of grammar in L2 teaching and learning has led to the introduction of various assessment materials to test grammatical knowledge and mastery of L2 learners. Thus, L2 academic writers, consciously or unconsciously, are taught to focus on grammar as one of the first requirements of a high quality piece of writing.

Secondly lexical mastery is basically thought of as a discipline-oriented skill or knowledge, while grammar is considered rather discipline-free. This has created more demand for structural or grammatical guidelines in nonnative speakers' (NNS) academic writing classes to cover a larger number of students from various disciplines, though lexical features are as much discipline-free as will be discussed below. Thus, this article was based on a motivation to contribute to the gap felt regarding considering lexical features in evaluating ESP academic writing texts.

2.0 LEXICAL FEATURES

Academic writing subgenres can be categorized as those written by students who are considered novice writers – ranging from simple class assignments to postgraduate theses – as well as those written by professional academic writers such as journal articles. Besides their differences, all types of written academic texts share some common features since all are firstly written, and secondly, academic. Hence, some similarities should be expected across all academic texts in terms of lexical, grammatical, and discourse features setting standards to evaluate the quality of an academic writing text as novice, professional, etc. Among various lexical features, three were investigated in this article: lexical density, complexity, and formality.

2.1 Lexical Density

Lexical density is a way to measure how closely the information is packed through words in a text. Lexical density as one of the distinguishing factors between written and spoken texts – written texts containing a higher lexical density – has been measured in different ways. One common way is the ratio of content (lexical) words to total words – content words referring to words such as 'nouns, verbs, adjectives, and adverbs' as opposed to function (grammatical) words which are 'prepositions, articles, pronouns, modals, auxiliary verbs, conjunctions, determiners, and particles' which function in grammatical system.

However, this model seems to have some limitations. In some cases, it may not be able to distinguish between the ways information is conveyed through words; as in the following two sentences.

S1. This figure illustrates how various environmental factors are getting improved currently.
S2. This figure is an illustration of various environmental factors current improvement.

Sentences 1 and 2 (S1 and S2) are two different surface structures of the same deep structure, which yield the same result if subjected to the above lexical density measurement. In both examples, which contain 11 words, 9 of which are lexical words, lexical density is $9 \div 11 = 0.82 \times 100 = 82$ %. As can be seen, lexical density defined as "the ratio of lexical words to total words" does not reveal any differences

here, while these two sentences are obviously different in the way information is conveyed through words.

Then, as another solution, we refer to Halliday's definition of lexical density. Halliday (1989:67) argues that information in text is packed in larger grammatical units – clauses – rather than words and suggests to measure lexical density along clauses as 'average number of lexical items per clause'. Halliday then offers the term 'lexical items' to be distinguished from lexical words. Lexical items, in this sense, are similar to lexical words but with the difference that some of them consist of more than one word. A lexical item can be defined as 'a series of content and function words which are connected to each other as a fixed expression and convey a special meaning and are used as a noun, a verb, an adjective, or an adverb', for instance, the expression 'at the moment' means 'now' and is used as an adverb.

Although lexical item is a meaning-based concept and thus useful in determining lexical density, this term is rather slippery and not as clear as lexical word and there might be cases of disagreement in counting the number of lexical items. Hence, in order to increase the reliability of the data, 'average number of lexical words per clause' seems to be more practical.

According to this definition, lexical density of sentences 1 and 2 will be different since the former is composed of two clauses containing nine lexical words, while the latter is composed of one clause containing nine lexical words. Then, lexical density of sentences 1 and 2 will be respectively 4.5 and 9, which sounds more appropriate and distinguishing.

The other advantage of this definition is that grammatical words are excluded which sounds more reasonable since they seem to be responsible for structural features rather than lexical features.

2.2 Lexical Complexity

Lexical complexity can be viewed from various aspects and for different purposes. One of the measurement methods of lexical complexity used in the literature is type/token ratio (TTR) (Gregg *et al.* 2002). TTR as defined by Biber *et al.* (1999) reveals vocabulary development throughout the text that is the result shows whether the writer has used a variety of words or has repeated the same words (for instance the word 'house' might be replaced with 'home' or 'residence'. However, it seems that it cannot be a good criterion for an academic text (for instance a postgraduate dissertation) since firstly the result varies with the length of the text and secondly an academic text is usually filled with words which are repeated many times which are of two nature; (a) either they are technical terms which are the focus of the study and hence are not interchangeable with any other synonyms, or (b) they are grammatical words such as 'articles, conjunctions, relative pronouns, etc.' which are again inevitable in the text. As an example, a part of a frequency list which was provided by Monoconc Software from the Introduction Chapter of a Master's dissertation is shown in Table 1.

Table 1 A part of the frequency list of a Master's Dissertation Introduction Chapter

N	Word	Freq.	%	Texts	%
1	THE	274	8.22	1	100.00
2	OF	165	4.95	1	100.00
3	TO	116	3.48	1	100.00
4	IN	103	3.09	1	100.00
5	AND	92	2.76	1	100.00
6	IS	73	2.19	1	100.00
7	A	58	1.74	1	100.00
8	ADJUDICATION	52	1.56	1	100.00
9	CONSTRUCTION	50	1.50	1	100.00

As can be seen, the most frequent words in this academic text are grammatical words followed by technical words which are the focus of the study and hence their repetition is unavoidable. Thus, TTR method does not reveal the true lexical complexity level in this text. Hence, the offered method for determining lexical complexity in this paper is based on meaning rather than form. Lexical complexity, especially in the field of translation, can be viewed from word meaning (Papi *et al*. 2007). Firstly, lexical complexity can be taken as polysemy or multiplicity of meaning of the same word in various contexts. Secondly, connotative meaning or additional meaning across different cultures is what adds to the complexity of word meaning in communication. Thirdly, abstractness of meaning makes comprehension more challenging and thus can be considered as an aspect of lexical complexity.

However, in NNS academic writing, none of these measures seem appropriate to assess lexical complexity of written academic texts. Firstly, in academic texts, a word is used in its most common meaning in a certain field of study, for instance the word 'complexity' in psychology and linguistics has two different meanings that are never found in the same text. Thus, polysemy may not be an appropriate measure for lexical complexity level. Secondly, academic words are not culture-bound since each term is of a clear definition which is realized and used by almost all academics in the same sense. Thirdly, the abstractness or concreteness of words in an academic text seems to depend on the nature of the discipline. For instance in psychology more abstract words are expected to be observed than in physics because psychology is an abstract science while physics is of a more concrete nature. Thus again this criterion cannot be an appropriate way to show the lexical complexity of the text. Hence, the following lexical complexity measurement is offered in this study.

Semantically speaking, lexical complexity can be defined as how a single lexical word is developed and how complicated a single lexical word would be to understand. This makes lexical complexity and density as two extremes of the same continuum: lexical density shows how information is packed through the combination of various words in context, while lexical complexity is about how the meaning is developed in a single lexical word out of context.

A closer look at academic words shows that a rather large part of academic terms are formed through adding Latin affixes (e.g. 'dis-', 'un-', 'de-') to word roots (e.g. power, hydrogen) to create terms such as 'disempowerment' or 'dehydrogenization'. As can be seen below, adding various affixes adds various meanings to the word roots.

Dis + em + **power** + ed = (a person) to be deprived of power or authority (e. g. Voters were disempowered by the new law)

Un + **power** + ed = not having or using power, specifically: not self-powered (e. g. an unpowered glider)

Then, as can be seen in Figure 1, it can be concluded that the more affixes are added to the word root the more complex the word meaning will be. Hence, the complexity of the words below increases downwards while all of them come from the same root 'power'.

Power
Em**power**
Disem**power**
Disem**power**ment

Figure 1 Relationship between the number of affixes and complexity

In English, affixes are divided into two types based on the changes they make on word meaning: derivational and inflectional (Brinton 2000, Akmajian 2001, and Yule 2006). Derivational affixes such as 're-, dis-, un-, -tion, -hood, etc.' make new words by adding concrete meaning to the words as in 'review, dislike, unsatisfied, collaboration, and childhood'. Inflectional affixes, on the other hand, change grammatical forms of the words not their meanings. All English prefixes and most suffixes are derivational except for the seven inflectional suffixes as follows:

- Two noun inflectional suffixes: Plural marker -s, Possessive marker 's,
- Three verb inflectional suffixes: Past participle markers -ed/-en, third person present singular marker -s, and progressive marker -ing,
- Two adjective inflectional suffixes: Comparative marker -er, Superlative marker -est

Another source of lexical complexity besides adding derivational affixes to word roots is using compound words defined by Richards and Schmidt (2002:98) as: "A combination of two or more words which function as a single word…. Compound words are written either as a single word (e.g. waterway), hyphenated words (e.g. self-government), or as two words (e.g. police station)." However, two-word compound words (e.g. police station) do not seem to reflect the lexical complexity – as defined in this research that is 'how the meaning of a single lexical word is developed and complicated to understand'. Thus, two-word compounds should be taken as two single words rather than one.

In this case, lexical complexity method offered in this study is based on the derivational affixes as well as compound words as will be explained below:

Firstly, we need to have a list of all words used in the text with their frequency. This can be done using various concordance software packages such as MonoConc. Then each word will be given a mark based on the number of its derivational affixes or compound parts. Simply, each word will get one mark for every derivational affix or compound part. In marking, we may come to a large number of words without any affixes or compound parts. These words can be divided into two categories: (a) those words that either can take affixes or can be combined with other words and thus have the potentiality to change into more complex forms, and (b) those words that cannot be used with any affixes or other words. The former group includes words such as word roots: add, move, etc. while the latter group is mainly composed of functional words such as: and, an, or, etc. We can give the score of 1 to the first group and the score of 0 to the second group to exclude them from the calculation since they do not play any roles in adding to the lexical complexity. Although some of these words may combine with

other words to form slangs, their combinability is considered only in an academic context. Finally, each number will be multiplied by the frequency of that word in the text.

The obtained number depends on the length of the text, in other words, longer texts contain more words and thus will have a higher lexical complexity. Thus this number should be divided by total number of the words in the text to reveal average lexical complexity of the text. Table 2 shows part of a Master's dissertation subjected to this measurement:

Table 2 Measuring lexical complexity of part of a Master's dissertation using the offered method in this study

Words	Frequency	Score	Frequency × score
A	9	0	0
Ability	2	2	4
About	3	0	0
Academic	1	2	2
According	1	1	1
Acquired	1	1	1
Add	1	1	1
Addition	2	2	4
Adjustments	1	2	2
Affect	1	1	1
Also	1	0	0
Among	1	0	0
An	2	0	0
And	20	0	0
Total	46	12	16

This text is composed of totally 46 words with a total lexical complexity of 16. Thus, average lexical complexity for this text will be 16÷46=0.34.

A glance at the following two sentences (S3 and S4) also approves this measurement. Sentences 3 and 4 are two different surface structures of the same deep structure, however, as it can be seen, in sentence 4, using derivational affixes has resulted in fewer words but with more complex meaning than sentence 3 (e. g. voters = the people who voted, reconsideration = to be considered again, etc). Table 3 illustrates lexical complexity of these two sentences.

S3: The people who voted wanted the new law to be considered again since it was not effective.
S4: Voters asked for reconsideration of the new law for its ineffectiveness.

Table 3 Lexical complexity of S3 and S4

S3				S4			
Words	**Frequency (fr)**	**Score (sc)**	**fr × sc**	**Words**	**Frequency (fr)**	**Score (sc)**	**fr × sc**
The	2	0	0	Voters	1	2	2
People	1	1	1	Asked	1	1	1
Who	1	0	0	For	2	0	0
Voted	1	1	1	Reconsideration	1	3	3
Wanted	1	1	1	Of	1	0	0
New	1	1	1	The	1	0	0
Law	1	1	1	New	1	1	1
To	1	0	0	Law	1	1	1
Be	1	0	0	Its	1	0	0
Considered	1	1	1	Ineffectiveness	1	4	4
Again	1	0	0				
Since	1	0	0				
It	1	0	0				
Was	1	0	0				
Not	1	0	0				
Effective	1	2	2				
Total	17	8	8	Total	11	12	12
Lexical Complexity= 8÷17=0.47				Lexical Complexity=12÷11=1.09			

2.3 Lexical Formality

As Goatly (2000) states English vocabulary can be divided into three groups based on formality; Old English words which are the most frequent ones occurring across various range of genres; French words – entered into English in the 14th Century – which are of medium formality; and Greek and Latin words – borrowed into English from the 16th to the 17th Century – which are the most cultivated, learned, and technical words and of the least frequency. For instance the words 'help, aid, and assist' refer to the same concept but with different formality and thus frequency levels. The word 'help' is the most frequent one; the word 'aid' is a French word and slightly formal; and the word 'assist' is a French word originally borrowed from Latin used as a technical term and thus the least frequent of all. As another instance, the words 'pad, house, and residence' refer to the same concept, however; the word 'pad' is a slang used by a few people; the word 'residence' is a technical term used in special contexts; while the word 'house' is the most common and thus found in various genres.

Thus, lexical formality is correlated with frequency. Highly informal words (slangs) or highly formal words (technical terms) are found in particular communities as a means of identifying their members, providing solidarity among them, and excluding outsiders such as a group of doctors discussing a patient's critical situation in his presence, or discussing a medical subject in a conference. Thus, based on the literature, Figure 2 is suggested in this article to illustrate the relationship between formality and frequency:

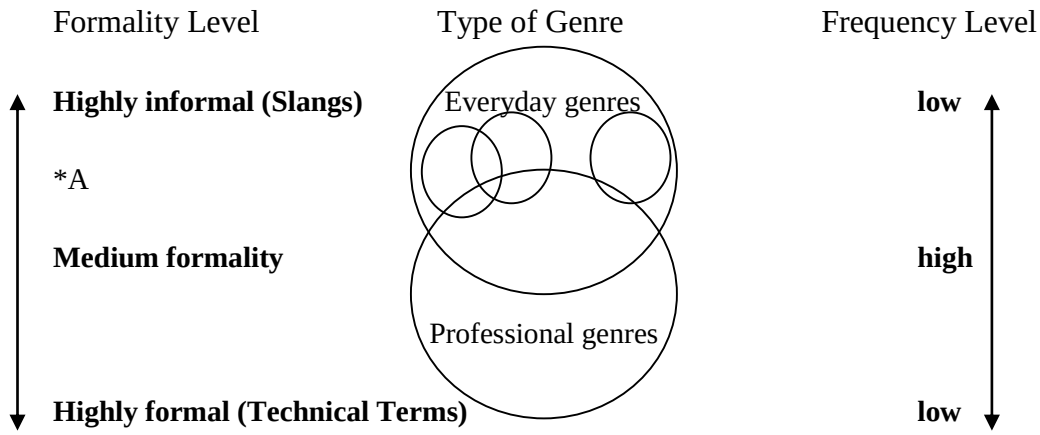

Figure 2 Formality and frequency

Figure 2 is composed of three columns: Formality level, Type of genre, and Frequency level. The first column, Formality Level, is a continuum from highly informal to highly formal words. In fact, there is no clear-cut border to separate two neighboring formality levels. For instance, the words in the position *A, fall between the two formality levels and might be classified as slangs by some linguists or as words of medium formality by others. However, formality levels are more distinguishable moving further along the continuum.

The second column, Type of Genre, shows how the words of various formalities are used across different genres. In this figure, genres are classified as two main types: everyday and professional. In each type there are many subgenres (smaller circles), which may or may not overlap each other. For instance, language used among high school students, criminals, or sportsmen are instances of everyday genre, while medical language, academic language, and law language are different professional genres. As illustrated in Figure 1, Professional genres seem to contain a large number of technical terms as well as words of medium formality but no slangs. Everyday genres, on the other hand, seem to contain a large number of slangs as well as words of medium formality but no technical terms. Thus, the words in both ends of the formality continuum are only used in certain genres but words in the middle of this continuum are used in a wider range of genres.

Finally, as depicted in the last column, highly informal and formal words are of lower frequency since they are only used in particular genres, while, words of medium formality are of more frequency because they are shared by more people and thus used in a wider range of genres. Frequency level is also a continuum which increases towards the center and decreases towards the ends and there is no clear-cut border between words of lower and higher frequencies.

Native speakers (NSs) of a language have a true understanding of lexical formality required in various social contexts; however, NNSs lack this knowledge. Many studies have been conducted on native speakers' and nonnative speakers' vocabulary knowledge to reveal the large gap between them. For instance, Nation and Waring (1997 cited in Hinkel 2004a) commented that a five-year-old NS child has a vocabulary range of 4000-5000 word families; an average university student 17000; and a university graduate student around 20000. While an adult NNS may have less than 5000 words (Hinkel

2004a). Based on various large corpora of English, many vocabulary lists and dictionaries considering frequency and formality have been created such as the Academic Word List (Coxhead 2000), Collins COBUILD dictionaries, etc.'

Measuring lexical formality of academic text is very challenging and no appropriate way has been suggested so far. To achieve this goal, we need to refer to language dictionaries, academic dictionaries, as well as academic wordlists to provide a list of the words that must be avoided in academic texts together with a list of recommended replacements for them based on their formality level. Then, using concordance software we will be able not only to assess the text based on the frequency of words with different formality levels but also to suggest better terms to replace the inappropriate ones.

3.0 CONCLUSION, IMPLICATION, AND SUGGESTIONS FOR FUTURE RESEARCH

Lexical features are as important as grammatical features and play as important a role in the quality of academic writing and deserve as much attention by the teachers, students, and methodologists. Besides grammatical features, lexical features also need to be observed in academic written texts, for instance, all types of academic written texts need to meet certain level of lexical density, complexity, formality, etc. However, determining the quality of various lexical features in academic writing texts requires more research since lexical features seem to be influenced by the nature of the academic disciplines as well as sentence structures they are used in. Thus, it is suggested that in the future more research be driven towards developing new methods for measuring various lexical features in academic texts. This way, a clear picture of the nature of professional lexical features can be obtained that will be useful in guiding the students in their academic writing achievements. Even as a result of the highly computable nature of lexical features, software packages such as concordance software packages can be designed to assess and evaluate lexical features of students' writing and to provide advice in terms of vocabulary choice.

REFERENCES

Akmajian, A., R. A. Demers, A. K. Farmer, and R. M. Harnish. 2001. *Linguistics: An Introduction to Language and Communication*. Fifth edition. USA: Massachusetts Institute of Technology.

Biber, D., Johansson S., Leech, G., Conrad, S., and E. Finegan. 1999. *Longman Grammar of Spoken and Written English*. London: Longman.

Billig, M. 2008. Nominalizing and De-Nominalizing: A Reply. *Discourse & Society*. 19/6: 829–841.

Brinton, L. J. 2000. *The Structure of Modern English: A Linguistic Introduction*. Amsterdam & Philadelphia: John Benjamins.

Coxhead, A. 2000. 'A New Academic Word List.' *TESOL Quarterly*. 34/2: 213–238.

Crawford, W.J. 2005. Verb Agreement and Disagreement: A Corpus Investigation of Concord Variation in Existential There + Be Constructions. *Journal of English Linguistics*. 33/1: 35–61.

Goatly, A. 2000. *Critical Reading and Writing: An Introductory Coursebook*. London: Routledge.

Gregg, N., Coleman, C., Stennett, R. B., and M. Davis. 2002. Discourse Complexity of College Writers With and Without Disabilities: A Multidimensional Analysis. *Journal of Learning Disabilities* 35/1: 23–38.

Halliday, M. A. K. 1989. *Spoken and Written Language*. Second edition. Oxford: Oxford University Press.

Hinkel, E. 2001. Giving Examples and Telling Stories in Academic Essays. *Issues in Applied Linguistics* 12/2: 149–70.

Hinkel, E. 2002. *Second Language Writers' Text: Linguistic and Rhetorical Features*. Mahwah, NJ: Lawrence Erlbaum Associates.

Hinkel, E. 2004a. *Teaching Academic ESL Writing: Practical Techniques in Vocabulary and Grammar*. Mahwah, NJ: Lawrence Erlbaum Associates.

Hinkel, E. 2004b. Tense, Aspect and the Passive Voice in L1 and L2 Academic Texts. *Language Teaching Research* 8 /1: 5–29

Hyland, K. 2008a. Academic Clusters: Text Patterning in Published and Postgraduate Writing. *International Journal of Applied Linguistics* 18: 41–62.

Hyland, K. 2008b. As Can Be Seen: Lexical Bundles and Disciplinary Variation. *English for Specific Purposes* 27: 4–21.

Master, P. 1991. Active Verbs with Inanimate Subjects in Scientific Prose. *English for Specific Purposes* 10/1: 15–33.

Nation, P. and R. Waring 1997. Vocabulary Size, Text Coverage, and Word Lists. In N. Schmitt and M. McCarthy (eds.). *Vocabulary: Description, Acquisition, and Pedagogy*. Cambridge: Cambridge University Press. 6–19.

Papi, M. Bertuccelli, G. Cappelli, and S. Masi (eds.). 2007. *Lexical Complexity: Theoretical Assessment and Translational Perspectives*. Pisa: Plus Pisa University Press.

Richards, J. C. and R. Schmidt. 2002. *Longman Dictionary of Language Teaching and Applied Linguistics*. Third edition. London: Longman (Pearson Education).

Samraj, B. 2002. Introductions in Research Articles: Variations across Disciplines. *English for Specific Purposes*. 21/1: 1–17.

Samraj, B. 2005. An exploration of a Genre Set: Research Article Abstracts and Introductions in Two Disciplines. *English for Specific Purposes*. 24/2: 141–156.

Samraj, B. 2008. A Discourse Analysis of Master's Theses across Disciplines with a Focus on Introductions. *English for Academic Purposes*. 7/1: 55–67.

Swales, J. M. 1990. *Genre Analysis: English in Academic and Research Settings*. Cambridge: Cambridge University Press.

Wei, Y., and L. Lei 2011. Lexical Bundles in the Academic Writing of Advanced Chinese EFL Learners. *RELC Journal* 42/2: 155–166.

Yule, G. 2006. *The Study of Language*. Third edition. Cambridge: Cambridge University Press.